DOCUMENT RESUME

ED 403 317 TM 026 053

AUTHOR Baker, Eva L.; And Others
TITLE CRESST: A Continuing Mission To Improve Educational Assessment. Evaluation Comment.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 96
CONTRACT R3053600002
NOTE 29p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Agenda Setting; Educational Assessment; Educational Policy; Educational Practices; *Educational Research; Elementary Secondary Education; *Evaluation Methods; *Information Dissemination; Partnerships in Education; *Program Development; *Research Design; Research Methodology; Test Bias; Test Use; Validity
IDENTIFIERS Center for Research on Eval Standards Stu Test CA; Large Scale Assessment

ABSTRACT

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) is a partnership of the University of California at Los Angeles, the University of Colorado at Boulder, Stanford University, The RAND Corporation, the University of Pittsburgh, the Educational Testing Service, and the University of California, Santa Barbara. This issue of "Evaluation Comment" shares the goals and perspectives that will shape CRESST's research program for the next 5 years. With a focus on the assessment of education quality, CRESST expects to study: (1) assessment that leads to improvement in teaching and learning; (2) understanding and influencing assessment policy and large-scale practice; (3) improved technical knowledge about the quality of assessment; and (4) dissemination and outreach that successfully decreases the interval between research and practice. The conceptual model that will underlie the research program emphasizes societal impact as the ultimate goal and identifies four major domains: validity, fairness, credibility, and utility. This model will guide an ambitious agenda of research focusing on the areas of system coherence, adaptations and accommodations of assessments, the measurement of progress, and reporting. The issue also discusses the CRESST conference scheduled for September 1996 and 1996 CRESST resource papers and technical reports. (Contains 1 figure and 137 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 403 317

UCLA's Center for the Study of Evaluation &
The National Center for Research on Evaluation, Standards, and Student Testing

EVALUATION COMMENT

Summer 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. C. BEER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

CRESST: A Continuing Mission to Improve Educational Assessment

Eva L. Baker, Robert L. Linn, and Joan L. Herman

The newly awarded National Center for Research on Evaluation, Standards, and Student Testing (CRESST) is a partnership of UCLA, the University of Colorado at Boulder, Stanford University, The RAND Corporation, the University of Pittsburgh, the Educational Testing Service, and the University of California, Santa Barbara. In this *Comment*, we share the goals and perspectives that will shape our research program for the next five years. With the assessment of education quality our focus, we commit ourselves to four key programs of work: assessment that leads to improvements in teaching and learning; understanding and influencing assessment policy and large-scale practice; improved technical knowledge about the quality of assessment; and dissemination and outreach that successfully decreases the interval between research and practice. Our programs are driven by our desire to meet the immediate and future needs of education policy and practice, yet reflect the historical lessons and current assessment trends across America.

Trends in Assessment Policy

Throughout this century educational testing has been called upon to serve many different purposes. It has been used to allocate scarce resources through student selection, to place children in educational programs, to monitor student achievement, and to hold educators accountable for student performance. Reformers have used test results to document deficiencies in order to help build the case that change was needed. They have also relied on testing as a major instrument of reform (see, for example, U.S. Congress, Office of Technology Assessment, 1992). Not surprisingly, testing has also been at the center of frequent and sometimes intense controversy (Cronbach, 1975).

ALSO IN THIS ISSUE...

The CRESST Conference ..... p. 2

New CRESST Resource Papers & Technical Reports ..... p. 25

(continued on page 4)

TM026053

BEST COPY AVAILABLE

The 1996 CRESST Conference

Please mark your calendars to attend the 1996 CRESST Conference at UCLA's Sunset Village, September 5-6, 1996. CRESST partners and other distinguished colleagues will present findings from recent K-12 assessment research and discuss issues for upcoming research projects. A tentative schedule of presenters and sessions is provided below and on pages 23 and 24.

The *on-site* $275 registration fee includes all meals and housing at the conference center for two days, a reception and formal dinner. Extra nights including meals are available at $95. The *commuter fee* for those not requiring housing is $125 and includes several meals, parking, and a reception.

We must receive your registration form by **August 19** and total payment by **September 3**. Space is limited to the first 300 registrants. Sorry, but absolutely no partial plans are available, and no refunds or changes may be made after September 3.

Look for additional conference details in the summer *CRESST Line* issue and at the CRESST Web site, http://www.cse.ucla.edu. Or call CRESST at 310-206-1532.

The 1996 CRESST Conference Agenda (Tentative)

Note: Presenters, sessions, and titles are subject to change.

Thursday, September 5, 1996

8:45-10:30 a.m. — **The CRESST Assessment Model: Consolidating What We Know and Where We Need to Go**

- Overview of Key Validity Issues — *Robert Linn*, CRESST/University of Colorado at Boulder
- Overview of Key Equity Issues — *Edmund T. Gordon*, CRESST/Yale University (Emeritus)
- Creating Credible Assessments for the Public — *Richard Colvin*, Los Angeles Times
- The Politics of Credibility — *Lorraine McDonnell*, CRESST/University of California, Santa Barbara

10:45-noon — **Validity and Utility of Assessment Systems**

- Standards for Assessment Systems — *Joe Conaty*, Office of Educational Research and Improvement (invited)
- Large-Scale Systems Serving Multiple Purposes: The Title I Standards and Assessment Challenge — *Eva L. Baker*, CRESST/UCLA
- One State's Response to the Challenge: The Washington State Example — *Judy Billings*, Washington State Dept. of Public Instruction
- The Face of Cultural Diversity in Assessment System Design — *Roland Tharp*, University of California, Santa Cruz

1:15-2:45 p.m. — **The CRESST Road Map: Priority R&D Issues in Reaching Our Destination**

- Framing the Future of Assessment Systems — *Joan Herman*, CRESST/UCLA
- From Standards to Assessments — *Thomas Romberg*, University of Wisconsin, Madison
- Measuring Student Progress — *Bengt Muthén*, CRESST/UCLA
- System Consequences for At-Risk Students — *George Madaus*, Boston College

3:00-4:00 p.m. — **Special Sessions From Centers' Recent Research**

- Alignment of Content Standards and Assessment Measures in Mathematics and Science — *Norman Webb*, Wisconsin Center for Education Research
- School and Classroom Interventions for At-Risk Students — *Sylvia Johnson*, Center for Research on the Education of Students Placed at Risk
- An Overview of Research From the National Assessment of Educational Progress — *Jamal Abedi*, CRESST/UCLA; George Bohrnstedt, American Institutes for Research
- Assessing Problem Solving in Science — *Noreen Webb*, CRESST/UCLA; *Gail Baxter*, CRESST/University of Michigan (invited)

(continued on page 23)

1996 CRESST CONFERENCE REGISTRATION FORM

Thursday, September 5 - Friday, September 6, 1996

ALL REGISTRANTS: *Complete this form and mail with payment to: CRESST/UCLA, 10920 Wilshire Blvd., #900, Los Angeles, CA 90024-6511, Attn: Kathryn Morrison. Call or fax registration information immediately to ensure your space at the conference. Phone: (310)206-1532. Fax: (310)825-3883. We strongly suggest you make a copy of this form for you records. Reservations are due by* ***August 19, 1996.***

Name (print) _____

Title _____

Organization (for name badge) _____

Address _____

City _____ State _____ Zip _____

Phone _____ Fax _____ E-mail _____

ALL CONFERENCE ATTENDEES, INCLUDING PRESENTERS, MUST SPECIFY BELOW THE NIGHTS THEY REQUIRE A ROOM.

I will need a room for the following nights:

❑ Tues 9/3 ❑ Wed 9/4 ❑ Thur 9/5 ❑ Fri 9/6 ❑ Sat 9/7 ❑ None (Off-Site Registrant)

PRESENTERS ONLY

Presenters' airline reservations **MUST** be made through American Express Travel at (800) 235-8252. Travel forms must be submitted after the conference for reimbursable expenses. **Specify housing above.**

Registration, meals, and room fees are waived for presenters.

Date and Time of Presentation: _____

Audio-Visual Needs (overhead projectors provided): _____

_____

UCLA FACULTY, STUDENTS & STAFF

Indicate if you are:

❑ UCLA Faculty

❑ UCLA Grad Student

❑ CRESST Research Staff

Fees are waived for UCLA faculty and CRESST staff.

ON-SITE OR COMMUTER REGISTRANTS ONLY

Registration options: On-site or Commuter. The $275 on-site conference registration fee includes Wednesday and Thursday night housing, parking if necessary and all meals for two days at the Sunset Village conference center. Extra nights ($95 per night) include all meals and housing. The $125 commuter registration fee includes several meals and parking but no housing. If you need housing, Sunset Village is strongly recommended.

❑ On-site ($275)
(Wednesday & Thursday night housing included)

❑ Commuter ($125)

Extra nights at $95 per night:

❑ Tues ❑ Fri ❑ Sat ❑ Other (specify) _____

FEES: Checks payable to: ***Regents of UC***

Registration Fee	$ _____
Extra Night/s	$ _____
Total	$ _____

CRESST: A Continuing Mission To Improve Educational Assessment

Many complaints about testing during the 1970s and 1980s emphasized bias against minority, female, and disadvantaged students (e.g., Haney, 1981; National Commission on Testing and Public Policy, 1990) and secrecy. More recently, the debate has highlighted the perceived mismatch between tests designed to measure general achievement without clear ties to specific curriculum or instructional experiences and the need for assessments that are explicitly linked to particular content standards or curriculum guidelines (Resnick & Resnick, 1992). The latter approach represents a shift in the formulation of the basic constructs to be measured by formal assessments from general ability to learned accomplishments and a desire to use the same assessment for different purposes.

This reformulation occurred for a number of reasons. First, traditional forms of assessment were gradually demystified. The 1979 test disclosure legislation in New York (S. B. 5200-A and subsequent amendments to Article 7-A of the New York Education Law) resulted in the publication of previously secure admissions tests and allowed leisurely perusal of items heretofore seen only in times of stress by respondents. Acknowledgments were made that test preparation could help performance (Bond, 1989; Messick & Jungeblut, 1981; Pike, 1978), especially when it led to generalized improvements in relevant knowledge and skills, for example, understanding of mathematics (Johnson & Wallace, 1989). Questions about norming practices—the Lake Wobegon effect—raised by Cannell (1987) and examined by technical experts (for example, Koretz, 1988; Linn, Graue, & Sanders, 1990, Shepard, 1990) brought public discussion to previously unchallenged procedures. Changing views of student learning (see, e.g., Brown & Campione, 1994; Chi, Glaser, & Farr, 1988; Glaser, 1996; Greeno, 1995) suggested different sorts of tests (Archibald & Newman, 1988; Baron, 1990; Frederiksen, 1984; Glaser & Silver, 1994; Mislevy, 1994; Shavelson, Lang, & Lewin, 1994; Stiggins, 1987; Wiggins, 1989) and led to the growing interest in performance assessment.

In the winter of 1991, OERI awarded an R&D center with a new assessment mission—a mission that focused on the design and validation of these new types of performance assessments and studies of the impact of these assessments in practice. In partnership with teacher organizations, the research community, Council of Chief State School Officers, numerous state leaders, district assessment personnel, and an extended cadre of classroom teachers, CRESST articulated its mission in a list of criteria for the validity of new assessments (Linn, Baker, & Dunbar, 1991).

Performance assessments suffered a crisis of credibility that continues today...

Almost simultaneously, a national movement began focusing on content standards and the idea of connecting assessments deeply to clear expectations (National Council on Education Standards and Testing, 1992; Smith & O'Day, 1991). Supporting legislation in state after state and the activity of professional and scientific organizations, such as the National Council of Teachers of Mathematics (1989a, 1989b) and the National Academy of Sciences (1993), created a sense that U.S. assessment practices would undergo a significant change. Compatible changes were also being sought for the assessment and certification of accomplished teachers. Yet, just as this enterprise gained momentum, reservations about these approaches also surfaced. Performance assessments suffered a crisis of credibility that continues today, a split that displays the

larger gap between the views of educational reformers and other segments of the public (Johnson & Immerwahr, 1994).

> ...opposition to a new performance-based test was based on propriety of assessment content, perceived objectivity, and cost of administration and scoring.

Some critics of new assessments objected to the idea that standards had a "national" rather than local inspiration (Bracey, 1995; Sizer, 1995). Contention about the content of some standards led to a reconsideration of the wisdom of a national approach (Brimelow & Spencer, 1995; Rich, 1995). The California experience is a case in point, where opposition to a new performance-based test was based on propriety of assessment content, perceived objectivity, and cost of administration and scoring (e.g., Asimow, 1994). Opponents argued that the test neglected fundamental skills and academic content and, heightened by rumors that the assessment asked students to write personal experiences, therefore invaded family privacy. CRESST interviews with parents in those schools where the opposition was highest suggest that lack of information and misunderstanding of the assessment contributed as much to parental concerns as did the content and new format of the test. In addition, analyses of California's and other new performance assessments showed deficiencies in some technical properties (Cronbach, 1995; Select Committee, 1994).

These objections, meritorious or otherwise, influenced the current state of assessment system development—a strategy far more complex than that envisioned in 1991, involving different formats of measures to meet public expectations. Yet, response to the credibility concern may be at the expense of the validity of system information. Many systems are adopting strategies that emphasize the local rather than national development of curriculum standards and expectations (Higuchi, 1995), more cautious advances on new forms of assessments, and a recommitment to standardized tests. Supported by the Improving America's Schools Act (1994), the use of multiple measures to meet expectations of different constituencies and a focus on the inclusion of all students in assessments are characteristics of these assessments. The inclusion and accommodation requirements signal a change in the definition of fairness—from the protection of subgroups to the exposure of any differences in their performance with the intent of stimulating improved system efforts to alleviate the revealed inequities.

> Pressures will mount in these new systems to combine, equate, or solve methodologically messy conceptual, and possibly intractable, conflicts.

These decisions create enormous challenges with regard to the formulation of approaches to study the quality of these information systems and the fairness, utility, and societal impact of the results they yield. We must recognize that any one element of one system will be subject to rapid change stemming from public perception, policy realignment, or from technical quality concerns. Pressures will mount in these new systems to combine, equate, or solve methodologically messy conceptual, and possibly intractable, conflicts. It is our intention to address these broad issues as much as possible in the real-time operations of states, districts and schools, rather than in the cleaner, neater world of leisurely reanalysis and occasional data collection. For it is only in these

settings that we will be forced to confront the reality of public perception and technical quality.

The current social and policy context leads us to a mission ultimately focused on the range of information in assessment systems. Although the specifics will vary greatly, there are a few enduring questions that apply to systems and to individual measures. Is the information produced credible? Are the resulting inferences supported? Does the assessment lead to desired actions? Is the testing useful for the different purposes it is intended to serve? Much of CRESST's R&D will be guided by these broad questions.

Conceptual Model for CRESST Research: Contributing to Knowledge, Educational Improvement, and Public Engagement

For assessment systems to benefit education, they must provide accurate information, they must be conceived as precursors to reflection and action, and they must address the multiple frames of references of their users—the public, teachers, administrators, policy makers, and, most of all, students. In Figure 1 we lay out the conceptual model underlying our new research program. It emphasizes societal impact as our ultimate goal: We seek to produce new knowledge and understanding about educational quality, to contribute to the use of assessment systems for educational improvement—both in policy and accountability uses and in teaching and learning—and to encourage productive, public engagement in education. The model identifies four major domains: validity, fairness, credibility, and utility. We assert that the utility and ultimate impact of assessment systems depend on the validity, fairness and credibility of the information produced by the system. All three characteristics of assessment are necessary. Assessments of high technical quality are of little use unless their results are credible to key audiences. Similarly, credible but invalid or unfair results will falter in the long run, for they will produce misleading interpretations or counterproductive, inequitable actions. Therefore, validity, fairness, credibility, and utility provide the conceptual framework for the upcoming CRESST research and development program.

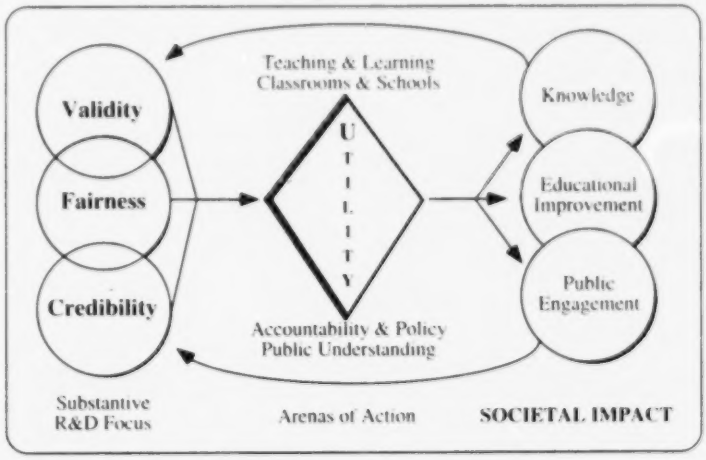

Figure 1. CRESST conceptual model.

Validity

Validity is the core technical concept in educational assessment (see, for example, AERA, APA, & NCME, 1985; Baker, O'Neil, & Linn, 1993; Cronbach, 1971, 1980, 1988; Linn, 1994; Messick, 1989, 1994; Shepard, 1993), and a comprehensive view of validity drives our work. In his authoritative chapter on validity, Messick (1989) defines validity as "an integrated, evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13).

> In simple language, construct validity is concerned with the meaning of the measures.

Our work particularly highlights issues of construct validity. In simple language, construct validity is concerned with the meaning of the measures. Writing, mathematics, and history achievement are examples of constructs; so, too, are personal attributes such as motivation or self-concept. We address construct validity by asking questions such as: Does this procedure, intended to measure problem solving, actually reflect higher order abilities rather than recall? Does this set of questions on geographic concepts provide a sound basis for generalizing about a class's understanding of the domain of geography? How much do difficulties with the English language impair the performance of some pupils on a mathematics assessment?

Construct validity for a certain purpose is undermined when the assessment is too narrow ("construct underrepresentation") and also when it is too broad ("construct-irrelevant influences"). For example, in a physics assessment, construct underrepresentation occurs if the assessment tasks require responses to only a few non-representative ideas rather than a proper sample of physics topics or if the topics are inappropriately weighted. Construct-irrelevant influences are defined in terms of ancillary skills, that is, skills other than those that the assessment is intended to measure, that influence performance (Haertel & Wiley 1993; Wiley & Haertel, 1996). Understanding of the language used in the assessment is perhaps the most common example of an ancillary skill where the intent of the assessment is knowledge of a content area. Other examples of ancillary skills include testwiseness, differential familiarity with task formats, and personal characteristics such as test anxiety or impulsivity.

Fairness

Fairness is an essential aspect of the validity of any assessment or accountability system. The ideal is that the quality of the inferences drawn from information will be judged in terms of appropriateness for all people, of all backgrounds and needs. For this ideal to be approached, fairness must pertain to every aspect of the assessment process, from assessment design, through administration, and interpretation of results. Fairness becomes increasingly important as the stakes attached to results are raised. The general perception that the system is fair is also central to the credibility of assessment results.

> One common view of fairness is objectivity, safeguards to assure that no one gets special advantage...

Although fairness in testing has universal appeal, differing conceptions of fairness have been applied to the measurement process and have different implications for both technical inferences and action. One

common view of fairness is objectivity, safeguards to assure that no one gets special advantage, an essential component of the underlying rationale for the standardization of test administration and scoring (Cronbach & Suppes, 1969). A second aspect of fairness, the avoidance of bias, has often been taken to imply the avoidance of disadvantage, but more technically is the goal ". . . to limit the differential validity of a given interpretation" (Cole & Moss, 1989, p. 205). In this conception of fairness, the construct to be measured is assumed to have the same meaning for all subgroups.

A substantial amount of inquiry in the area of bias has focused on identifying test content and features with negative impact for subgroups (Cole & Nitko, 1981; Figueroa & Garcia, 1994), addressing questions such as the following: Does the assessment put minority students at a disadvantage because they are less familiar than their majority group counterparts with content that is not essential to the skills and understandings that the assessment is intended to measure? (See, for example, Johnson, 1995.) Does the task context or any other non-essential aspect give an unfair advantage to boys in comparison to girls, or vice versa? (See Tittle, 1975.) Do skills that are ancillary to the intent of the assessment, such as reading ability for a mathematics problem-solving assessment, create an unfair disadvantage for certain students, for example, students with limited English proficiency? (See August, Hakuta, & Pompa, 1994; Haertel & Linn, 1996; Haertel & Wiley, 1993; Wiley & Haertel, 1996.) Does the assessment reflect situations or problems that are likely to be culturally biased? (See Gordon, 1992; Winfield, 1995.) Potential bias with the scoring of performance tasks must also be considered (Baker & O'Neil, 1996).

A third conception of fairness has a more compensatory and active flavor, focusing on the adaptations and accommodations in the assessment process that would provide students with an opportunity to display their competence. Regardless of conceptions, one point is clear: Available analytic techniques for examining the fairness of assessments, for example, linguistic complexity of tasks (e.g., Abedi, Lord, & Plummer, 1994), differential item functioning (e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993), or sensitivity reviews, neither provide a guarantee of fairness nor are sufficient to support the overall validity judgment. Taking seriously the concept of fairness may well lead to alternative pathways, including subgoals and measures, to attain common standards or to achieve more diverse goals for all students.

Credibility

I have very simple questions. Are the schools getting better or worse? What can I do to help? I can't get a straight answer.

—Chair of State Legislative Education Committee 1994

The third feature of our model, credibility, encompasses the perceptions and values of the public in general and of participants directly involved in education activities and decisions. Unfortunately, at the present time many Americans do not feel well served by educational information they receive; they feel closed out and uninformed (McDonnell, 1995). The diverse and shifting positions on controversial educational issues have increased suspicion about the validity, veracity, and propriety of centralized sources of information. In addition to privacy concerns, a significant segment of the public, often fed by cynicism of many in the media, believe that they are not being given a truthful picture by arms of government at the district, state, or national levels (Gunther, 1992).

Validity without credibility produces assessments that have no lifespan and whose findings are con-

tested, diminished, or dismissed. Validity, objectivity, and fairness are elements that influence the degrees of trust placed in findings by different constituencies. Moreover, credibility depends on the quality of information, the way results are communicated, and the purposes and uses to which results are put. Although credibility has not typically been examined by those who study assessment, it has been the focus of research attempting to understand how the public uses information in forming opinions about political issues (e.g., Zaller, 1992), how they judge the trustworthiness of the media (e.g., Gaziano & McGrath, 1986; Gunther, 1992; West, 1994), and how policy elites and the public regard the veracity of policy analysis and of various social indicators such as unemployment statistics (Bozeman, 1986; Innes, 1990; MacRae, 1985).

Utility

How can I use this in my teaching?

—10th-grade history teacher, 1994

The domain of utility addresses constraints and options for action in the real world. We distinguish three main components: potential utility, action, and impact.

- By *potential utility* we refer to the fit between the assessment, its design and intended purposes, and practical constraints for its use, including user perceptions. Potential utility depends on whether assessments are coherent, feasible, and cost sensitive, and whether purposes and results are clear and understandable to users.
- By *action* we mean the degree to which and how assessments are actually translated into practices and policies: how the assessment is actually used. Action depends on professional development for users, supportive resources, and implementation strategies.
- By *impact* we mean the degree to which desired effects are produced and unintended negative effects are avoided. The model conceives impact as identifiable, shorter-term consequences for practice and policy, particularly the fairness accruing to all parties. Broader, less direct impact of information includes effects on public engagement in and public views on educational matters and longer term changes in policy and educational systems. These broader, ultimate goals are conceived as societal impact in our model.

The potential utility of new and traditional forms of assessment has been widely debated (Herman, 1992; Resnick & Resnick, 1992; Shepard, 1995; Wiggins, 1989; Wolf, Bixby, Glenn, & Gardner, 1991), and substantial research has identified the general challenges to moving from potential to the reality of action and impact (Gearhart & Wolf, in press; Herman & Klein, in press; Koretz, Stecher, Klein, & McCaffrey, 1994; Stecher & Herman, in press).

Utility depends on validity, fairness and credibility...

The utility of an assessment or accountability system may vary for different actors and audiences. Three prime audiences—teachers, policy makers, and the general public—are emphasized in the proposed research and development work of CRESST. Our arenas of action are assessment systems for teaching and learning and large-scale assessment systems for equity, policy, and public understanding. Utility depends on validity, fairness, and credibility but may be as much influenced by local circumstances, unforeseen expectations of constituencies, and the personal predilections of leadership.

CRESST: A Continuing Mission To Improve Educational Assessment

The heart of our model of utility is the person, the human dimension, not abstract methodology, a particular analytic technique or any preferred form or format of test. For it is, after all, people who must make inferences that are accurate, fair, and appropriate for particular purposes and students. People make the judgments about whether they can trust, can understand, and will value and use information. People take action or do nothing; their choices of action fit to real limits of available knowledge, sense of benefit, and understanding of costs. Creating better methods, high-precision techniques, more inclusive assessments, and glossier, high-tech reports of results is of little use unless people use assessment results wisely to achieve worthy goals.

Core Problems Guiding the New CRESST Research Agenda: Assessment System Goals and Validity Agenda

Given the sheer number of issues that arise in the validity, fairness, credibility, utility and ultimate impact of any assessment system, we have chosen to highlight here four research areas in our new program: system coherence, adaptations and accommodations of assessments, the measurement of progress, and reporting.

System Coherence and Multiple Measures

Heavy demands are placed on assessments. They are expected to serve a range of purposes, yet their ability to do so requires coherence within and across elements of the educational system as well as within the system's assessments (Smith & Levin, 1996).

Aligning assessment and curriculum. Both past experience and common sense indicate that assessments do more than simply provide information when assessment results are made highly visible and used to hold educators accountable. Assessments influence what gets emphasized in the classroom and what falls by the wayside (Koretz, 1988; Madaus, 1988; Shepard, 1991; Smith, 1991). Indeed, it is the recognition that assessments can influence instruction that contributes to their appeal to policy makers as potential tools of educational reform.

One of the lessons learned from accountability systems of the past is that system coherence is essential. If the assessments that count for accountability purposes are out of alignment with desired classroom practice, they will reshape the enacted curriculum to mirror the accountability measures and, if inappropriate, distort teaching and learning. In contrast, where clearly connected to important system goals, assessments may support desirable coherence. Not only must assessments be aligned with content and performance standards; the whole system—curriculum materials, teaching strategies, professional development, incentives, sanctions, and expectations of various levels—must be aligned.

A given assessment may be well suited to meet some expectations but poorly adapted for others.

Connecting the information from multiple measures. Another lesson learned from accountability assessments of the past is that it is difficult for a single assessment to serve multiple purposes well without a major redesign effort (Baker, Linn, Abedi, & Niemi, 1996). A given assessment may be well suited to meet some expectations but poorly adapted for others. A teacher, for example, needs specific information on an immediate basis to guide short-term instructional decisions. External assessments yield information that is both too general and too slow in coming to be useful for making day-to-day instructional decisions. The informal assessment information that teachers rely upon for those day-to-day and

moment-to-moment decisions, on the other hand, even when compiled in a student portfolio (Gearhart & Herman, 1995), is too idiosyncratic to be useful for informing policy makers or the public objectively about overall student achievement.

> Such [test] proliferation obviously increases the overall assessment burden...

The various, often incompatible, demands placed on assessments have led to recommendations that assessments be tailored to specific uses. Although nominally sensible, such recommendations, in turn, have led to a proliferation of testing. This is evident from even a simple listing of the assessments in which a given student may be required to participate—teacher-made assessments, instructionally embedded tests that accompany textbooks and instructional materials, criterion-referenced tests required by the district, a norm-referenced test used for program evaluation, a criterion-referenced statewide assessment, and even, perhaps, the Trial State Assessment of the National Assessment of Educational Progress. Such proliferation obviously increases the overall assessment burden and, concomitantly, creates a problem of integrating findings into sensible inferences and actions. Unfortunately, proliferation does not necessarily solve the problem of matching assessments to use, since several of the assessments may still be expected to serve the same purpose, and other purposes may be inadequately served. Moreover, multiple assessments can lead to real and apparent conflicts in both interpretation and use of results when the different assessments emphasize different content or types of skills.

In any event, assessment and accountability systems almost always involve multiple measures. At the simplest level, these may simply be scores in different content areas. More complicated systems may include multiple types of assessment data such as teacher-scored student portfolios, results of centrally-scored performance assessment tasks, and standardized tests. Although each measure may yield useful results when considered alone, an assessment system implies the combination of information in meaningful ways. Sometimes, actions are required by legislation or board policy based on cumulative information across all measures, for instance, the decision to designate a school for a school improvement program or the distribution of awards to schools (see, for example, Crone, Long, Franklin, & Halbrook, 1994; Improving America's Schools Act of 1994; Kentucky Department of Education, 1995; Mandeville, 1988; Sanders & Horn, 1993).

> ...procedures and strategies are needed to assure that the multiple measures contribute to, rather than undermine, coherence.

Systems clearly need to allow for the flexible inclusion of multiple measures, but procedures and strategies are needed to assure that the multiple measures contribute to, rather than undermine, coherence. In particular, there are two potential problems that need to be addressed in dealing with multiple measures: (a) Redundancy across measures in the underlying constructs assessed may introduce unintended weightings in composite or summary scores; and (b) Important distinctions may be lost when results of multiple measures are combined. Multivariate analyses are needed to disentangle the redundancy and expose the different aspects of performance that support the overall validity of system interpretations. An important focus of CRESST

research, system coherence also requires that the information from multiple measures be combined in ways that are consistent with purposes and that provide information about status and progress.

Adaptations and Accommodations

Recent federal legislation (Improving America's Schools Act, 1994; Individuals with Disabilities Education Act, 1990) presents states, districts and schools with new challenges in providing disabled students the least restrictive environment and in encouraging states and districts to include in their large-scale assessments students with Individualized Education Plans (IEPs) and language minority students who have traditionally been excluded. How do we accommodate the needs of students of varying abilities and disabilities within mainstream instructional and assessment programs? How do we adapt assessments to the needs of language minority students whose achievement has heretofore been largely unexamined? How do we assure accurate placement of students with varying abilities and language capabilities? There is little research to date to guide policy and practice (August et al., 1994). Particularly perplexing is the issue of assuring fairness to all students—both those who are designated for special services and accommodations and those who are not.

Problems of exclusion. A common practice in the past has been simply to excuse students from participation in the assessment for whom it is deemed inappropriate. Yet such exclusions raise important fairness issues and can distort overall assessment results (Haladyna, Nolen, & Haas, 1991), particularly when the rules for exclusion vary from one site to another or from one time to another at a given site. Moreover, without inclusion, the assessment system provides no information about a sometimes sizable proportion of the students participating in the education system, which may reduce the likelihood that these students receive the services they need to achieve the content standards being assessed.

Recent experience with NAEP provides some indication of the scope of the exclusion problem. In the 1992 NAEP administration, approximately 5% of the sampled students were excluded from the assessment because an IEP judged it inappropriate for the student to participate in the assessment. The State Trials showed that IEP exclusions ranged from 2% to 8% (National Academy of Education, 1993), a range that likely reflects differences in state and local policies and practices rather than any differences in special learning needs of students from state to state. On the other hand, variations in exclusion rates for language minority students, which ranged from less than 2% to 11% for Grade 4 reading, likely reflect differences in immigration patterns as well as differences in inclusion policies and practices.

The pressure is to exclude students—from the assessment and, perhaps, from meaningful instruction.

Though seldom discussed in public reports of results, exclusion rates on district administered standardized tests often are as high or higher than those on NAEP, and many state assessments also exclude a substantial number of students because of language minority or IEP status. Exclusions are generally motivated by concern that the assessment is inappropriate for the excluded student, but in some high-stakes situations, exclusion from testing or retention may also come about as a way of inflating scores (Darling-Hammond, 1995; Zlatos, 1994). The pressure is to exclude students—from the assessment and, perhaps, from meaningful instruction.

Issues in inclusion. Clearly assessments designed for large-scale, on-demand administrations may be

BEST COPY AVAILABLE

inappropriate for some students due to language requirements or because the tasks and response demands are not suitable for the current instructional levels of disabled students (Thurlow, Ysseldyke, & Silverstein, 1993). Special needs students who are to be held to the same standards as other students may need accommodation in test format, for instance, large-print versions of the test, or in testing environment, for example, a test carrel (Amos, 1980; Beattie, Grise, & Algozzine, 1983; Wildemuth, 1983). Other accommodations for students with disabilities may also include more breaks during testing, or extended testing time, perhaps over several days. Inclusion of students who have been excluded in the past clearly cannot achieve the goal of fairness unless participation of those students is meaningful and leads to valid interpretations and actions (i.e., the assessment interpretations and uses have an acceptable degree of construct validity) (Sherman & Robinson, 1982).

> ...there has been little or no research investigating the validity of inferences from these adaptations or alternatives.

In spite of available or newly developed adaptations, there are some students for whom these test adaptations are still inappropriate. Alternative assessments are needed for these students (see Kentucky Portfolios for Special Education, Kentucky Department of Education, 1995). Although promising, there has been little or no research investigating the validity of inferences from these adaptations or alternatives.

Some states have made a strong commitment to the idea of including all or nearly all students by offering assessments in languages other than English and by allowing for adaptations or accommodations for students with diverse needs, for example, extra time, or oral administration (see Kentucky Department of Education, 1995), but have not examined the construct equivalence of these measures. Spanish language versions of assessments in content areas have been used in some states (e.g., California and Texas), simply ignoring the confounding issues of language of instruction, prior educational history, and cultural differences (Durán, 1992; Geisinger, 1992; Valdés & Figueroa, 1994). Offering a test in both English and Spanish, furthermore, does not assess subject area competency of students not fully literate in either language.

> Defensible uses and interpretations of results based on adaptations and accommodations need to be articulated and justified.

Parents and teachers of students with disabilities and those with limited English language proficiency want their children included in testing for purposes of accountability and to improve the education of their children. At the same time, they certainly do not want their children hurt, frustrated, or treated unfairly by inclusion. At issue, then, is how eligibility for accommodations and adaptations should be determined, what modifications should be permitted, and how scores obtained under nonstandard conditions should be reported. Defensible uses and interpretations of results based on adaptations and accommodations need to be articulated and justified. Inappropriate uses and interpretations also need to be identified.

Choices about the basis for accommodation are influenced by empirical evidence concerning the equivalency of constructs measured in different lan-

guages (LaCelle-Peterson & Rivera, 1994), background knowledge (Cole & Scribner, 1973; Johnson, 1992), instructional opportunity (Baker & Rogosa, 1995; Herman, Klein, Heath & Wakai, 1994; Pullin, 1994; Winfield & Woodard, 1993), and motivation (Ogbu, 1978). It may be unfair, for example, to use student performance on an assessment aligned with newly adopted content standards and curriculum to compare or make quality judgments about teachers who have had differential access to professional development activities designed to introduce the standards and curriculum. Similarly, it may be unfair to compare or judge students based on their performance on assessments that are consistent with the instruction provided to some students but out of alignment with that provided to other students. Thus, an analysis of the correspondence between what is taught and what is assessed will be an important aspect of the CRESST agenda. How validity studies can best deal directly with alignment issues is yet to be determined.

> ...validation research is essential to provide both the evidential and consequential basis to support specific adaptations and accommodations and interpretations of results that they yield.

As was argued by the National Academy of Science panel on Placing Children in Special Education: A Strategy for Equity (Heller, Holtzman, & Messick, 1982), validation research is essential to provide both the evidential and consequential basis to support specific adaptations and accommodations and interpretations of results that they yield. For this reason, the CRESST project focusing on adaptations and accommodations for language minority students and the study for IEP students outline research and development plans within a broad validation framework, and classroom-based projects address the problem from the perspective of actual teaching and learning issues.

> Parents, students, teachers, administrators, policy makers, and the public share interest in a simple question: Are [my] children making progress?

Measuring Progress

Learning involves change. Hence it is no surprise that the measurement of change is of fundamental interest to many assessment and accountability systems. Parents, students, teachers, administrators, policy makers, and the public share interest in a simple question: Are [my] children making progress? Are they learning? Are schools getting better? New Title I regulations, furthermore, create a basic and substantial need to measure change in terms of students' "annual yearly progress."

Yet the measurement of change poses substantial challenges, including problems of low reliability of change scores, confounding changes in what is measured with changes in student performance, and sensitivity of growth to the particular type of scale used to report assessment results. Other problems arise when the goal is the assessment of progress for groups (e.g., the identification of schools that are making adequate progress), most notably the potential confounding effect of changes in the student population due to year-to-year differences or due to mobility.

Sensitivity of measures of change to the scale of measurement is also a cause for concern because of the arbitrary nature of scales often used to report results of assessments. With standardized tests, for example, the pattern of growth in student achievement appears quite different for scores based on

different scaling models (Linn, 1981; Linn & Slinde, 1977; Seltzer, Frank, & Bryk, 1994). Performance-based assessments have not been studied as extensively with regard to this issue, but they are also subject to the problem that change results are sensitive to choice of scale. Regardless of form of assessment, the use of standards-based reporting procedures raises yet other complications, since the changes reported for individuals and for groups of students will be sensitive not only to gains in student achievement, but to the number and stringency of standards used, as well as where on the scale the standard is set.

Assessment and accountability systems clearly need to be capable of reporting progress as well as status of schools and districts, including intermediate benchmarks that can be used to gauge the adequacy of the progress. This principle implies the need to attend to several technical issues, such as the comparability of assessments from year to year and, in the case of schools or school systems, the comparability of different cohorts of students. Two of the more important issues that need to be dealt with in the proposed CRESST research are the development of adequate procedures for estimating the degree of uncertainty associated with measures of student and school progress, and effective communication of that information to audiences that will use measures of progress.

Fortunately, there have been substantial improvements in the analytical approaches now available for tackling the problems associated with measuring progress. New analytic models and perspectives on the measurement of change (e.g., Bryk & Raudenbush, 1987, 1992; Muthén, Huang, Jo, Khoo, Nelson Goff, Novak, & Shih, 1995; Muthén, Khoo, & Goff, 1994; Rogosa, Brandt, & Zimowski, 1982; Rogosa & Saner, 1995a, 1995b; Rogosa & Willett, 1985) provide a firmer theoretical foundation for attacking the problems associated with the demand to measure student progress and the progress of educational systems.

...research is needed to provide a basis for understanding the implications of using different summaries of student performance...

Analytical models described by Bryk and Raudenbush, Rogosa, and by Muthén and his colleagues will serve as the starting point for CRESST research and development work on progress measurement. While value-added conceptions provide a useful framework for addressing many of the goals implicit in the demand to report the progress of schools or other aggregations of students, substantial research and development is needed to understand how best to deal with student mobility and to understand the implications and trade-offs of models that rely on year-to-year comparisons of different cohorts of students enrolled in a given grade (for example, Grade 4 students in 1996-97 compared to Grade 4 students in 1997-98) as compared to approaches that rely on longitudinal samples following the same students across years. Similarly, research is needed to provide a basis for understanding the implications of using different summaries of student performance, such as group means or percentage of students meeting a standard, for measuring progress.

Reporting for Understanding and Action

Without effective communication and reporting, the utility of an assessment or any assessment system is severely compromised: Results languish unused, the potential of substantial investments wasted; or worse, results can be misused. The proliferation of assessments nationally and locally and the addition of new forms of assessment only heighten historic

problems in teachers', students' and public understanding of test results and what to do with them (Hambleton & Slater, 1995; Herman & Dorr-Bremme, 1983; Stiggins, 1991). Recent media reports of the public response to new forms of assessment further underscore inherent problems of understanding and communication (Merl, 1994; Sanders, 1995). People—students, parents, teachers, administrators, policy makers, the public—cannot act sensibly on information they either do not have or do not adequately understand. The issues thus are ones of access and distribution as well as of the clarity and usability of information.

Just as a single test score cannot serve all purposes, a single report cannot meet the needs of all users.

Just as a single test score cannot serve all purposes, a single report cannot meet the needs of all users. As most students of writing know, the writer is supposed to develop a model of what the audience expects in level of information, tone, and structure. Successful writers create good matches with their audiences. Some excellent writers can adapt their work for a wide range of audiences differing in expectation, knowledge, language, tolerance for detail, desire for entertainment, and available time to devote to the enterprise. Similarly, research shows that users want reports tailored to their needs and decision arenas, with direct implications for action (Herman, 1989; Hood at al., 1972). Furthermore, research on multiple intelligences (Gardner, 1993) and other aspects of cognition suggests that multiple modalities of communication are essential to meet diverse cognitive styles (Snow & Lohman, 1989).

Technology provides new possibilities for displaying and customizing information and new avenues for distribution (Baker, in press). New iconic representations are possible to help guide users' attention and understandings. Desktop publishing and automated authoring and editing systems will greatly ease the burden of adapting user-friendly reports for different constituents, and automated analysis routines will enable different levels of data aggregation for different reports. While technology can ease the production problem, the specifics of what different audiences want to know, what they will find *credible*, and how best to communicate also demand attention and remain prerequisite issues that will be addressed by our research programs. In collaboration with relevant constituencies, we will seek to understand how to combine and communicate complex information from assessment systems in ways that are fair, valid, and credible for different audiences and to serve different purposes.

As the Internet demonstrates, a major shift in information access is underway; distributed use of information, tools, and systems is now a reality.

The power of an interactive communication process to promote information use also is well established (Patton, 1988). In this area, too, technology dramatically opens up channels for users' interaction and analysis. As the Internet demonstrates, a major shift in information access is underway; distributed use of information, tools, and systems is now a reality. Wider distribution to new users will require clear frameworks (Baker & O'Neil, 1994) for interpreting results. Users of information will want to know how it relates to other findings. Parents and teachers will become interested in how they can replicate what is assessed (create local versions of standard measures), use new approaches to help their own children succeed, and discuss and improve

the educational process. Our research program will help to identify the requirements of credible and useful information systems for these various users and to build tools to support their use of assessment.

Addressing an Ambitious Agenda

CRESST has established an ambitious agenda for the next five years. The specific work we have proposed is guided by a shared set of beliefs about the nature of effective R&D in our mission area:

- Assessment, evaluation, and accountability represent only a small part of what is truly important about educating our children.
- R&D must commit to improving educational quality.
- Useful R&D focuses on real problems and uses theoretical paths to explore their solution.
- Collaboration is essential in identifying and understanding problems and in determining the value of options.
- R&D findings should be aggressively communicated in accessible and compelling ways to all audiences—policy makers, politicians, teachers, parents, and students. Don't wait for them to ask.
- Diverse perspectives are needed to clarify real differences and to find equitable, workable balances.
- Impartiality, not advocacy, is the key to the credibility of research and development.
- The best R&D meets current needs but seeks to redefine constraints so that creative solutions are possible.

We address our mission through four highly interrelated programs:

- Program One–*Assessment in Action* addresses fundamental problems in improving the utility of assessment at the school and classroom levels and, in the process, will investigate substantive issues of validity, fairness, and credibility that are essential in assessment systems serving large-scale accountability and educational improvement purposes.
- Program Two–*Accountability, Equity, Policy, and Public Engagement* combines active, scholarly reflection on the purposes, implementation, and effects of large-scale assessment systems with action-oriented, practical responses to current assessment design and interpretation problems.
- Program Three–*Technical and Functional Quality of Assessment Systems: Validity, Equity, and Utility* investigates the technical uses and interpretations that are made from assessment results and the changes produced in the school, community, and student body; in curriculum and instruction and policies for student advancement and placement; in policies and attitudes of the district or community; and in the natural perception of education.
- Program Four–*Outreach and Dissemination* will access the continuous feedback necessary for the improvement of CRESST R&D and greatly reduce the cycle of time between assessment research and its application to practice.

Our programs in teaching and learning and in accountability, equity, and policy reflect the arenas of action for our research. We consider them points of entry that over time will enable us to design assessment systems that can serve both arenas. Our program organization mirrors our belief that the improvement of assessment policy and practice requires sophisticated understanding of socio-political, functional, and technical problems of assessment systems as they exist from the top down (Program Two) as well as the intricacies of how assessment is

used and perceived from the bottom up, at the school and classroom learning levels (Program One). Merging these two perspectives with the theoretical advances in fairness and validity (Program Three), we believe that over the next five years we can make an important contribution to the design and analysis of coherent assessment systems that serve educational quality.

References

Abedi, J., Lord, C., & Plummer, J. R. (1994). *Language background as a variable in NAEP mathematics performance* (Draft Technical Report). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Amos, K. M. (1980). Competency testing: Will the LD student be included? *Exceptional Children, 47*(3), 194-197.

Archibald, D. A., & Newman, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in secondary schools*. Washington, DC: National Association of Secondary School Principals.

Asimow, N. (1994, February 25). Alice Walker story furor grows. *San Francisco Chronicle*, p. A20.

August, D., Hakuta, K., & Pompa, D. (1994). For all students: Limited English proficient students and Goals 2000. *Occasional Papers in Bilingual Education, 10*, 4.

Baker, E. L. (in press). Readying NAEP to meet the future. In E. Bohrnstedt (Ed.), *Evaluation report on the 1994 NAEP Trial State Assessment*. Palo Alto, CA: National Academy of Education.

Baker, E. L., Linn, R. L., Abedi, J., & Niemi, D. (1996). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research, 89*(4), 197-205.

Baker, E. L., & O'Neil, H. F., Jr. (1994). (Eds.). *Technology assessment in education and training*. Hillsdale, NJ: Lawrence Earlbaum.

Baker, E. L., & O'Neil, H. F., Jr. (1996). Performance assessment and equity. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises, problems, and challenges* (pp. 183-199). Mahwah, NJ: Lawrence Erlbaum Associates.

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*(12), 1210-1218.

Baker, E. L., & Rogosa, D. (1995, April). *Sleepless in Woodland Hills: The Leigh Burstein legacy*. Presentation at the annual meeting of the American Educational Research Association, San Francisco, CA.

Baron, J. B. (1990). Performance assessment: Blurring the edges among assessment, curriculum, and instruction. In A. B. Champagne, B. E. Lovitts, & B. J. Calinger (Eds.), *Assessment in the service of instruction*. Washington, DC: American Association for the Advancement of Science.

Beattie, S., Grise, P., & Algozzine, B. (1983). Effects of test modifications on the minimum competency performance of learning disabled students. *Learning Disability Quarterly, 6*(1), 71-77.

Bond, L. (1989). The effects of special preparation on measures of scholastic aptitude. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 429-444). New York: Macmillan.

Bozeman, B. (1986). The credibility of policy analysis: Between method and use. *Policy Studies Journal, 14*, 519-539.

Bracey, G. W. (1995). *Final exam: A study of the perpetual scrutiny of American education*. Bloomington, IN: Technos.

Brimelow, P., & Spencer, L. C. (1995). Comeuppance. *Forbes, 155*(4), 121-128.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229-270). Cambridge, MA: MIT Press.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147-158.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cannell, J. J. (1987). *National-normed elementary achievement testing in America's public schools: How all 50 states are above the national average* (2nd ed.). Daniels, WV: Friends of Education.

BEST COPY AVAILABLE

CRESST: A Continuing Mission To Improve Educational Assessment

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, M., & Scribner, S. (1973). Cognitive consequences of formal and informal education. *Science, 182*, 553-559.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-219). New York: Macmillan.

Cole, N. S., & Nitko, A. J. (1981). Measuring program effects. In R. A. Berk (Ed.), *Educational evaluation methodology: The state of the art*. Baltimore, MD: Johns Hopkins University Press.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30*, 1-14.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In *New directions in tests and measurements, No. 5*. San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1995). *A valedictory: Reflections on 60 years of educational testing. Board Bulletin*. Washington, DC: National Research Council, Board on Testing and Assessment. (Available through National Information Service)

Cronbach, L. J., & Suppes, P. (1969). *Research for tomorrow's schools: Disciplined inquiry for education*. Stanford, CA: National Academy of Education/Macmillan.

Crone, L. J., Long, Franklin, B. J., & Halbrook, A. M. (1994). Composite versus component scores: Consistency of school effectiveness classification. *Applied Measurement in Education, 7*, 303-321.

Darling-Hammond, L. (1995). Equity issues in performance-based assessment. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 89-114). Boston, MA: Kluwer Academic.

Durán, R. P. (1992). Clinical assessment of instructional performance in cooperative learning. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 137-156). Washington, DC: American Psychological Association.

Figueroa, R. A., & Garcia, E. (1994). Issues in testing students from cultural and linguistically diverse backgrounds. *Multicultural Education, 2*(1), 10-19.

Fredericksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193-202.

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York: Basic Books.

Gaziano, C., & McGrath, K. (1986). Measuring the concept of credibility. *Journalism Quarterly, 63*(3), 451-462.

Gearhart, M., & Herman, J. L. (1995, Winter). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability. *Evaluation Comment*. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., & Wolf, S. A. (in press). Issues in portfolio assessment: Assessing processes from their products. *Educational Assessment*.

Geisinger, K. F. (1992). Fairness and selected psychometric issues in the psychological testing of Hispanics. In K. F. Geisinger (Ed.), *Psychological testing of Hispanics* (pp. 17-42). Washington, DC: American Psychological Association.

Glaser, R. (1996, April). *Learning, instruction, and assessment: A research agenda for the future*. Presentation at the annual meeting of the American Educational Research Association, New York.

Glaser, R., & Silver, E. (1994). *Assessment, testing, and instruction: Retrospect and prospect*. (CSE Tech. Rep. No. 379). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Gordon, E. W. (1992). *Implications of diversity in human characteristics for authentic assessment* (CSE Tech. Rep. No. 341). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Greeno, J. G. (1995). Understanding concepts in activity. In C. A. Weaver III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 65-95). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gunther, A. C. (1992). Biased press or biased public: Attitudes toward media coverage of social groups. *Public Opinion Quarterly, 56*, 147-167.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (NCES Rep. No. 96-802). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

BEST COPY AVAILABLE

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359-384). Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

Hambleton, R. K., & Slater, S. C. (1995). *Are NAEP executive summary reports understandable to policy makers and educators?* (Final Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Haney, W. (1981). Validity, vaudeville and values: A short history of social concerns over standardized testing. *American Psychologist, 36*(10), 1021-1034.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity. Panel on Selection and Placement of Students in Programs for the Mentally Retarded, Committee on Child Development Research and Public Policy, Commission on Behavioral and Social Sciences and Education, National Research Council.* Washington, DC: National Academy Press.

Herman, J. L. (1989). *Data for effective decision-making* (CSE Tech. Rep. No. 298). Los Angeles: University of California, Center for the Study of Evaluation.

Herman, J. L. (1992). What research tells us about good assessment. *Educational Leadership, 49*(8), 74-78.

Herman, J. L., & Dorr-Bremme, D. (1983). Uses of testing in the schools: A national profile. In W. Hathaway (Ed.), *Testing in the schools: New directions for testing and measurement* (No. 19, pp. 7-17). San Francisco, CA: Jossey-Bass.

Herman, J. L., & Klein, D. C. D. (in press). Assessing equity in alternative assessment: An illustration of opportunity-to-learn issues. *Journal of Educational Research.*

Herman, J. L., Klein, D. C. D., Heath, T. M., & Wakai, S. T. (1994). *A first look: Are claims for alternative assessment holding up?* (CSE Tech. Rep. No. 391). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Higuchi, C. (Ed.). (1995, July). *Language arts content standards.* Los Angeles, University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hood, P. D., et al. (1972). *The educational information market study.* San Francisco: Far West Laboratory for Education Research and Development.

Improving America's Schools Act of 1994, Conference Report 103-761. Regarding Public Law 103-382, signed October 20, 1994, (pp. 6-33). Washington, DC: House of Representatives.

Individuals with Disabilities Education Act (IDEA), 1990.

Innes, J. E. (1990). *Knowledge and public policy* (2nd ed.). New Brunswick, NJ: Transaction.

Johnson, J., & Immerwahr, J. (1994). *First things first: What Americans expect from the public schools.* New York: Public Agenda.

Johnson, S. T. (1992). Extra-school factors in achievement, attainment, and aspiration among junior and senior high school-age African American youth. *Journal of Negro Education, 61*, 99-119.

Johnson, S. T. (1995). Visions of equity in national assessment. In M. T. Nettles & A. L. Nettles, (Eds.), *Equity and excellence in educational testing and assessment* (pp. 343-366). Boston: Kluwer Academic Publishers.

Johnson, S. T. & Wallace, M. B. (1989). Characteristics of SAT quantitative items showing improvement after coaching among Black students from low-income families: An exploratory study. *Journal of Educational Measurement, 26*, 133-145.

Kentucky Department of Education. (1995). *KIRIS Biennium I Technical Manual.* Frankfort, KY: Author.

Koretz, D. M. (1988). Arriving in Lake Wobegon. Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*(2), 8-15, 46.

Koretz, D. M., Stecher, B. M., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications, *Educational Measurement: Issues and Practice, 13*(3), 5-16.

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55-75.

Linn, R. L. (1981). Measuring pretest-posttest performance changes. In R. A. Berk (Ed.), *Educational evaluation methodology: The state of the art* (pp. 84-109). Baltimore, MD: Johns Hopkins University Press.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

CRESST: A Continuing Mission To Improve Educational Assessment

Linn, R. L., Baker, E. L., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5-14.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre and posttesting periods. *Review of Educational Research, 47*, 121-150.

MacRae, D., Jr. (1985). *Policy indicators*. Chapel Hill: University of North Carolina.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum. Eighty-seventh yearbook of the National Society for the Study of Education, Part I* (pp. 83-121). Chicago: University of Chicago Press.

Mandeville, G. K. (1988). School effectiveness indices revisited: Cross-year stability. *Journal of Educational Measurement, 25*, 349-356.

McDonnell, L. M. (1995, September). *Defining curriculum standards: The promise and the limitations of performance assessment in schooling*. Paper prepared for the conference "Efficiency and Equity in Education Policy," convened by the National Board of Employment, Education, and Training and the Centre for Economic Policy Research. The Australian National University, Canberra.

Merl, J. (1994, May 6). Furor continues to build over state's CLAS exams. *Los Angeles Times*, p. A1, 18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191-216.

Mislevy, R. (1994). *Test theory reconceived* (CSE Tech. Rep. No. 376). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Muthén, B., Huang, L. C., Jo, B., Khoo, S. T., Nelson Goff, G., Novak, J., & Shih, J. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis, 17*(3), 371-403.

Muthén, B., Khoo, S., & Goff, G. N. (1994). *Multidimensional description of subgroup differences in mathematics data from the 1992 National Assessment of Educational Progress* (Technical Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

National Academy of Education. (1993). *The Trial State Assessment: Prospects and realities. The third report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1992 Trial State Assessment*. Stanford, CA: Stanford University, National Academy of Education.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (1993) *Science, technology and the federal government: National goals for a new era*. Washington, DC: National Academy Press.

National Board for Professional Teaching Standards. (1989). *Toward high and rigorous standards for the teaching profession*. Detroit, MI: Author.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America*. Chestnut Hill, MA: Author.

National Council of Teachers of Mathematics. (1989a). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics (1989b). *Curriculum standards for teaching mathematics*. Reston, VA: Author.

National Council on Education Standards and Testing. (1992). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. Washington DC: Author.

Ogbu, J. (1978). *Minority education and caste*. San Diego: Academic Press.

Patton, M. (1988). *Utilization-focused evaluation* (2nd ed.). Newbury Park, CA: Sage.

Pike, L. W. (1978). *Short-term instruction, testwiseness, and the Scholastic Aptitude Test: A literature review with research recommendations* (CB RDR 77-78, No. 2; ETS Research Rep. No. 78-2). Princeton, NJ: Educational Testing Service.

Pullin, D. C. (1994). Learning to work: The impact of curriculum and assessment standards on educational opportunity. *Harvard Educational Review, 64*(1), 31-54.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston, MA: Kluwer Academic Publishers.

BEST COPY AVAILABLE

CRESST: A Continuing Mission To Improve Educational Assessment

Rich, F. (1995, January 26). Eating their offspring. *New York Times*, v. 144, pp. A19(N), A21(L).

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726-748.

Rogosa, D. R., & Saner, H. M. (1995a). Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics, 20*, 149-170.

Rogosa, D. R., & Saner, H. M. (1995b). Reply to discussants: Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics, 20*, 234-238.

Rogosa, D., & Willett, B. (1985). Understanding correlates of change by modeling individual growth. *Psychometrika, 50*, 203-228.

Sanders, P. (1995, September). *Review of the Kentucky Instructional Results Information System*. Presentation at the CRESST conference "Assessment at the Crossroads," Los Angeles, University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Sanders, W. L., & Horn, S. (1993). *An overview of the Tennessee Value-Added System (TVAAS)*. Knoxville: University of Tennessee.

Select Committee. (1994). *Sampling and statistical procedures used in the California Learning Assessment System*. Sacramento, CA: California State Department of Education (Reprinted in Cronbach, 1995).

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement in choice of metric. *Educational Evaluation and Policy Analysis, 16*, 41-49.

Shavelson, R., Lang, H., & Lewin, B. (1994). *On concept maps as potential authentic assessments in science*. (CSE Tech. Rep. No. 388). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*(3), 15-22.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73*(3), 232-238.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education. Evaluation of National Assessment of Educational Progress achievement levels. *Proceedings from the Joint Conference on Standard Setting for Large-Scale Assessments*. Washington, DC: National Center for Education Statistics.

Sherman, S. W., & Robinson, N. M. (1982). *Ability testing of handicapped people: Dilemma for government, science, and the public*. Washington, DC: National Academy Press.

Sizer, T. R. (1995). Silences. *Daedalus, 124*(4), 77-84.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Smith, M. S., & Levin, J. (1996). Coherence, assessment and challenging context. In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. National Society for the Study of Education, 95th yearbook*. Chicago, IL: National Society for the Study of Education/University of Chicago Press.

Smith, M. S., & O'Day, J. A. (1991) Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233-267). New York: Falmer Press.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Stecher, B., & Herman, J. (in press). Portfolios for large-scale assessment. In G. Phye (Ed.), *Handbook of educational assessment*. San Francisco: Jossey-Bass.

Stiggins, R. J. (1987). The design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33-39.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*, 534-539.

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). *Testing accommodations for students with disabilities: A review of the literature* (Synthesis Rep. No. 4). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tittle, C. K. (1975) Fairness in educational achievement testing. *Education and Urban Society, 8*, 86-103.

U.S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

West, M. D. (1994). Validating a scale for the measurement of credibility: A covariance structure modeling approach. *Journalism Quarterly, 71*(1), 159-168.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

CRESST: A Continuing Mission...

Wildemuth, B. M. (1983). *Minimum competency testing and the handicapped*. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation. (ERIC Document No. ED 289 886)

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessments: Promises, problems, challenges*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Winfield, L. F. (1995). Performance-based assessments: Contributor or detractor to equity? In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 221-241). Boston: Kluwer.

Winfield, L. F., & Woodard, M. D. (1993). Assessment, equity and diversity in reforming America's schools. *Educational Policy, 8*(1), 3-27.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of educational research* (Vol. 17, pp. 31-74). Washington, DC: American Educational Research Association.

Zaller, J. R. (1992). *The nature and origins of mass opinion*. New York: Cambridge University Press.

Zlatos, W. (1994, November 6). Scores that don't add up. *New York Times*, v. 144, Sec 4A, pp. ED28(N), ED28(L).

UCLA's Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing

Eva L. Baker, Co-director
Robert L. Linn, Co-director
Joan L. Herman, Associate Director
Pamela Aschbacher, Assistant Director
Ronald Dietel, Editor
Katharine Fry, Editorial Assistant

The work reported in this publication was supported under the Educational Research and Development Center Program PR/Award Number R305B600002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this publication do not reflect the position or policies of the National Insitute on Student Achievement, the Office of Educational Research and Improvement or the U.S. Department of Education.

CRESST Conference...*from page 2*

- Assessment and Instruction in Elementary Mathematics: What We've Learned — *Maryl Gearhart*, CRESST/UCLA; *Megan Franke*, CRESST/UCLA
- Model-Based Large-Scale Assessment — *David Niemi*, University of Missouri; *Zenaida Munoz*, CRESST/UCLA
- Linking Language Arts Standards to Assessments — *Charlotte Higuchi*, CRESST/Los Angeles Unified School District/United Teachers, Los Angeles

4:00-5:15 p.m. — **High Technology Applications for the Assessment of Student Knowledge and Learning**

- Lessons from CRESST Technology Research Programs — *Harry F. O'Neil, Jr.*, CRESST/University of Southern California
- Model-Based Computer Assessment of Problem-Solving Skills in Science — *Ron Stevens*, CRESST/UCLA
- Recent Developments in Computer Assessment at the Educational Testing Service — *Randy Bennett*, Educational Testing Service
- Discussants — *Eva L. Baker*, CRESST/UCLA and *David Rogosa*, CRESST/Stanford University

Friday, September 6, 1996

8:30-10:15 a.m. — **Helping All Students to Arrive Part I: Adaptations for Students With Disabilities**

- Assessment Policy and Practice Recommendations for Students With Disabilities — *James Ysseldyke*, University of Minnesota
- Validity Issues in the Assessment of Students With Disabilities — *Daniel Koretz*, CRESST/RAND
- Representative from a state department of education (to be determined)

10:30-noon — **Helping All Students to Arrive Part II: Accommodations for Language Minority Students**

- A Research Framework for Investigating Accommodations for Language Minority Students —

The 1996 CRESST Conference...*from page 23*

Lorrie Shepard, CRESST/University of Colorado at Boulder

- Linguistic Issues in the Assessment of Language Minority Students — *Lily Wong Fillmore*, University of California, Berkeley (invited)
- Accommodations in Cooperative Learning Scenarios for Language Minority Students — *Richard Duran*, CRESST/University of California, Santa Barbara
- A Framework for Equitable Assessment Policies for English Language Learners — *Charlene Rivera*, George Washington University

1:15-2:45 p.m. **Policy and Assessment Forums: Making It Happen Through Collaboration, Public Engagement, and Action**

The following sessions will be concurrent policy forums focusing on current assessment issues with ample time for questions and answers.

Assessment, Policy, and Public Engagement

- *Michael Cohen*, U.S. Department of Education
- *Michael Feuer*, National Research Council
- *Mark Slavkin*, Los Angeles City Board of Education
- *Leah Lievrouw*, CRESST/UCLA

New Directions in Statewide Assessment

- *Duncan MacQuarrie*, Washington Department of Education
- *Doris Redfield*, Department of Education, Commonwealth of Virginia
- *Wayne Martin*, Colorado State Department of Education
- *Brian Stecher*, CRESST/The RAND Corporation

Building Teacher Capacity for Improved Classroom Assessment

- *Hilda Borko*, CRESST/University of Colorado at Boulder
- *Lynn Winters*, Long Beach Unified School District

Special Issues in the Assessment of At-Risk Students in Large Urban Schools

- *Sidney Thompson*, Los Angeles Unified School District
- *Carole Perlman*, Chicago Public Schools
- *Ruben Carriedo*, San Diego School District

Other invited forum participants include:

- *David Stevenson*, U.S. Department of Education
- *Adrienne Bailey*, Council of the Great City Schools
- *Carl Cohn*, Long Beach Unified School District (invited)

Theresa Dozier, U.S. Department of Education (invited)

3:00-4:15 p.m. **Moving Ahead**

- From Learning Theory to Assessment Practice — *Lauren Resnick*, CRESST/University of Pittsburgh (invited)
- From Vision to Capacity: Building Teacher Understandings in Standards and Assessments — *Marilyn Monahan*, National Education Association (invited)
- From Disjunct to Convergence: Moving Toward Reality in Policy and Practice — *Pascal Forgione*, National Center for Education Statistics
- The Road Ahead — *Lee Shulman*, Stanford University (invited)

4:15-4:45 p.m. **Wrapping It All Up**

- *Eva L. Baker*, CRESST/UCLA
- *Robert L. Linn*, CRESST/University of Colorado at Boulder

New CSE/CRESST Technical Reports

Assessing the Validity of the National Assessment of Educational Progress: NAEP Technical Review Panel White Paper

Robert L. Linn, Daniel Koretz, and Eva L. Baker

CSE Technical Report 416, 1996 ($5.00)

Under a contract from the National Center for Education Statistics, the CRESST Technical Review Panel has conducted a series of research studies addressing the uses and interpretations of the National Assessment of Educational Progress (NAEP), oftentimes known as the nation's report card. This report summarizes the most important findings including the quality of NAEP data, the number and character of NAEP scales, the robustness of NAEP trend lines, the trustworthiness of and interpretation of group comparisons, the validity of interpretations of NAEP anchor points and achievement levels, the effects of student motivation on performance, the adequacy of NAEP data on student background and instructional experiences, and what is understood from NAEP reports by educators and policy makers.

Performance Puzzles: Issues in Measuring Capabilities and Certifying Accomplishments

Lauren Resnick

CSE Technical Report 415, 1996 ($5.50)

In this report, CRESST/University of Pittsburgh researcher Lauren Resnick explores major issues in using assessments as a means of defining standards and encouraging efforts to meet them. She discusses the differences between the purposes for traditional and newer types of assessment, issues of scoring reliability, generalizability of observed performance, and content and construct validity involving performance assessment and portfolios.

Evidence and Inference in Educational Assessment

Robert Mislevy

CSE Technical Report 414, 1996 ($5.50)

"Data" from educational assessments become "evidence" only with respect to conjectures about students and their work, says Robert Mislevy in this report based on his 1994 presidential address to the Psychometric Society. Those conjectures are constructed around notions of the character and acquisition of knowledge and skill, and shaped by the purpose of the assessment and the nature of the inference required. Using a detailed analytic framework, the author demonstrates how the concepts and tools of mathematical probability can help explain relationships between evidence and inference about students' knowledge, learning, and accomplishments.

The Role of Probability-Based Inference in an Intelligent Tutoring System

Robert Mislevy and Drew Gitomer

CSE Technical Report 413, 1996 ($5.50)

Probability-based inference in complex networks of interdependent variables is an active topic in statistical research, spurred by such diverse applications as forecasting, troubleshooting, and medical diagnosis. Based on an instructional tutoring system for learning to troubleshoot a military F-15 aircraft hydraulics system, the authors in this study explore the role of Bayesian inference networks for updating student models in intelligent tutoring systems (ITSs).

Latent Variable Modeling of Longitudinal and Multilevel Data

Bengt Muthén

CSE Technical Report 412, 1996 ($3.50)

This report gives an overview of some aspects of latent variable modeling in the context of growth and clustered data. The author emphasizes the

New CSE/CRESST Technical Reports

benefits that can be gained from multilevel as opposed to conventional modeling techniques that ignore the multilevel data structure. Large-scale educational surveys are used to illustrate key points.

A Simple Approach to Inference in Covariance Structure Modeling With Missing Data: Bayesian Analysis

Bengt Muthén

CSE Technical Report 411, 1996 ($2.50)

In this report, CRESST/UCLA researcher Bengt Muthén investigates an improved approach for educational analyses where there are significant amounts of missing data. The author found that a Bayesian approach developed by himself and Gerhard Arminger, offers a promising technique for missing data covariance structure modeling. The technique should soon be available in covariance structure software.

Issues in Portfolio Assessment: The Scorability of Narrative Collections

John R. Novak, Joan L. Herman, and Maryl Gearhart

CSE Technical Report 410, 1996 ($4.50)

This report provides a model for examining technical questions concerning the validity and reliability of large-scale portfolio assessment scores. One of the key findings was that the holistic scale of the CRESST "Writing What You Read" narrative rubric—a rubric designed to enhance teachers' understandings of narrative and to inform instruction—could be used reliably and meaningfully in large-scale assessment of narrative collections.

Final Report: Perceived Effects of the Maryland School Performance Assessment Program

Daniel Koretz, Karen Mitchell, Sheila Barron, and Sarah Keith

CSE Technical Report 409, 1996 ($5.50)

In this study, CRESST/RAND researchers investigated the effects of the Maryland School Performance Assessment Program (MSPAP) by surveying Maryland teachers and principals. General support for MSPAP as an instrument of reform (in contrast to its role as an assessment) was widespread among surveyed educators, but teachers' views of MSPAP as an assessment were mixed. Large majorities of both teachers and principals reported that MSPAP has been at least somewhat successful in meeting its goal of improving instruction.

Teachers reported relying on diverse methods to prepare students for MSPAP, ranging from broad improvements in instruction to narrowly focused test preparation, such as use of practice tests. Their explanations of MSPAP score gains in their own schools, however, raise the possibility that initial gains were inflated. About half of the surveyed teachers reported that work with practice tests and familiarity with the assessment had contributed a great deal to their gains, while only 15% to 20% said the same of improvements in knowledge and skills. The report recommends several lines of research to explore issues raised by these survey findings.

Many CSE/CRESST Reports may be downloaded from the CRESST Web Site at www.cse.ucla.edu.

Recent CSE/CRESST Technical Reports

Estimating the Costs of Student Assessment in North Carolina and Kentucky: A State-Level Analysis
Lawrence O. Picus, Alisha Tralli, and Suzanne Tacheny
CSE Technical Report 408, 1995 ($4.00)

Opportunity-to-Learn Effects on Achievement: Analytical Aspects
Bengt Muthén, Li-Chiao Huan, Sick-Toon Khoo, Ginger Nelson Goff, John Novak, and Jeff Shih
CSE Technical Report 407, 1995 ($2.50)

Teachers' and Students' Roles in Large-Scale Portfolio Assessment: Providing Evidence of Competency With the Purposes and Processes of Writing
Maryl Gearhart and Shelby Wolf
CSE Technical Report 406, 1995 ($4.00)

Patterns of Performance Across Different Types of Items Measuring Knowledge of Ohm's Law
Brenda Sugrue, Rosa Valdes, Jonah Schlackman, and Noreen Webb
CSE Technical Report 405, 1995 ($2.50)

Using Group Collaboration as a Window Into Students' Cognitive Processes
Noreen Webb, Kariane Nemer, Alexander Chizhik, and Brenda Sugrue
CSE Technical Report 404, 1995 ($2.50)

Instructional Influences on Content Area Explanations and Representational Knowledge: Evidence for the Construct Validity of Measures of Principled Understanding—Mathematics
David Niemi
CSE Technical Report 403, 1995 ($8.00)

Monitoring and Improving a Portfolio Assessment System
Carol Myford and Robert Mislevy
CSE Technical Report 402, 1995 ($4.50)

Comparing Reliability Indices Obtained by Different Approaches for Performance Assessments
Jamal Abedi, Eva Baker, and Howard Herl
CSE Technical Report 401, 1995 ($2.50)

Portfolio Driven Reform: Vermont Teachers' Understanding of Mathematical Problem Solving and Related Changes in Classroom Practice
Brian Stecher and Karen Mitchell
CSE Technical Report 400, 1995 ($5.00)

Measurement of Teamwork Processes Using Computer Simulation
Harold F. O'Neil, Jr., Gregory K. Chung, and Richard S. Brown
CSE Technical Report 399, 1995, ($5.00)

Cognitive Analysis of a Science Performance Assessment
Gail Baxter, Anastasia Elder, and Robert Glaser
CSE Technical Report 398, 1995 ($5.00)

Contact Kim Hurst at 310-206-1532 or "kim@cse.ucla.edu" for a current CRESST Product Catalog with additional listings.

Order Form

Attach additional sheet if more room is needed.

CSE/CRESST Products

CSE Number	Title	Number of Copies	Price per Copy	Total Price

POSTAGE & HANDLING
(Special 4th Class Book Rate)

Subtotal of	$0 to $10	add $1.50
	$10 to $20	add $2.50
	$20 to $50	add $3.50
	over $50	add 10% of Subtotal

ORDER SUBTOTAL _____

POSTAGE & HANDLING (scale at left) _____

California residents add 8.25% _____

TOTAL _____

Orders of less than $10.00 must be prepaid

Your name & mailing address—please print or type:

_____

_____

_____

❑ Payment enclosed ❑ Please bill me

❑ I would like to receive free copies of the *CRESST Line* and *Evaluation Comment* publications.

UCLA Center for the Study of Evaluation
1320 Moore Hall/Mailbox 951522
Los Angeles, California 90095-1522

ADDRESS CORRECTION REQUESTED (ED 63)

NONPROFIT ORG.
U.S. POSTAGE
PAID
U.C.L.A.

BEST COPY AVAILABLE

Erwin Flaxman
Director
Institute for Urban and Minority Ed./ERIC
525 West 120th St., Box 40
New York NY 10027